ANTUSD: A Large Chinese Sentiment Dictionary

Shih-Ming Wang and Lun-Wei Ku

中央研究院 資訊科學研究所
SINCE 1982
Institute of Information Science, Academia Sinica

LREC 2016

Outline

Motivation

Corpus Building

- Related Corpora
- CopeOpi
- Extended-HowNet (E-HowNet)

Demonstrative Experiment

- Preprocessing
- Features
- Results

Conclusion

Motivation

Sentiment Dictionary

- A building block of sentiment analysis & opinion mining
- Applied as markers or machine learning features

Motivation

Sentiment Dictionary

- A building block of sentiment analysis & opinion mining
- Applied as markers or machine learning features

Augmented NTU Sentiment Dictionary (ANTUSD)

- Lack of Chinese resource
- Big & complete
- Expert labeled sentiment & machine predicted sentiment scores

Related Corpora I

- Words and labels were collected from several sentiment corpora (2006∼2010)

Related Corpora I

- Words and labels were collected from several sentiment corpora (2006∼2010)

Word-base, context free

- NTUSD
- ACIBiMA

Related Corpora I

- Words and labels were collected from several sentiment corpora (2006∼2010)

Word-base, context free

- NTUSD
  - A widely used Chinese sentiment dictionary
- ACIBiMA
  - Built to test Chinese morphological structure and sentiment

Related Corpora I

- Words and labels were collected from several sentiment corpora (2006∼2010)

Word-base, context free

- NTUSD
  - A widely used Chinese sentiment dictionary
  - Labels: **POS** and **NEG** (2812/8276)
- ACIBiMA
  - Built to test Chinese morphological structure and sentiment
  - Labels: **POS**, **NEU**, **NEG**, **NONOP**, and **NOT**
  - **NONOP** consists of regular non-emotion words
  - **NOT** consists of incorrectly segmented words

Related Corpora II

Sentence-based, context dependent

- NTCIR Multilingual Opinion Analysis Test Dataset
- Chinese Opinion Tree Bank

Related Corpora II

Sentence-based, context dependent

- NTCIR Multilingual Opinion Analysis Test Dataset
  - Dataset for international opinion analysis contest (6, 7 and 8th NTCIR)
- Chinese Opinion Tree Bank
  - Incorporate syntactic information (Chinese Treebank) into sentiment analysis

Related Corpora II

Sentence-based, context dependent

- NTCIR Multilingual Opinion Analysis Test Dataset
  - Dataset for international opinion analysis contest (6, 7 and 8th NTCIR)
- Chinese Opinion Tree Bank
  - Incorporate syntactic information (Chinese Treebank) into sentiment analysis

Properties

- Label process: sentence → sentiment words
- Each word might have conflicting labels
- Labels: **POS**, **NEU**, and **NEG**
- Context information not included in ANTUSD

CopeOpi

Machine predicted sentiment score

- CopeOpi: A Chinese opinion-analysis system
- Sentiment scores of documents, sentences, words, and characters
- Polarity score of each character is calculated statistically
- Word by summing up characters; sentence by summing up words...

Extended-HowNet (E-HowNet)

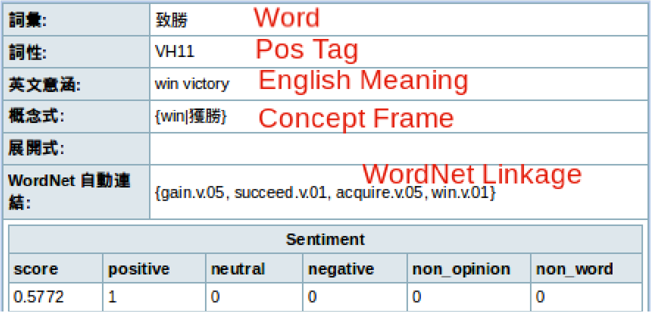
E-HowNet

- A frame-based entity-relation model extended from HowNet
- Define lexical senses (concepts) in a hierarchical manner
- Now integrated with ANTUSD and covers 47.7% words in ANTUSD

Extended-HowNet (E-HowNet)

E-HowNet

- A frame-based entity-relation model extended from HowNet
- Define lexical senses (concepts) in a hierarchical manner
- Now integrated with ANTUSD and covers 47.7% words in ANTUSD



詞彙:	致勝
詞性:	VH11
英文意涵:	win victory
概念式:	{win|獲勝}
展開式:	
WordNet 自動連結:	{gain.v.05, succeed.v.01, acquire.v.05, win.v.01}

Sentiment					
score	positive	neutral	negative	non_opinion	non_word
0.5772	1	0	0	0	0

Word

Pos Tag

English Meaning

Concept Frame

WordNet Linkage

Demonstrative Experiment

Experiment Setting

- Dataset: ANTUSD $\cap$ E-hownet, a total 12995 words
- Classifier: support vector machine (SVM) with linear kernel
- Average over 10-fold validation scores

Demonstrative Experiment

Experiment Setting

- Dataset: ANTUSD ∩ E-hownet, a total 12995 words
- Classifier: support vector machine (SVM) with linear kernel
- Average over 10-fold validation scores

Three sentiment analysis tasks

- Opinion extraction: identify opinion words ({**POS**,**NEG**} v.s. **NONOP**)
- Polarity classification: classify opinion words (**POS** v.s. **NEG**)
- Combined tasks (**POS**, **NEG**, **NONOP**)
  - $P = \frac{correct(opinion) \cap correct(polarity)}{proposed(opinion)}$
  - $R = \frac{correct(opinion) \cap correct(polarity)}{gold(opinion)}$
  - $F - score = \frac{2PR}{P+R}$

Preprocessing

Extract single label for each word

1. **NOT**: Count(Not)>0
2. **NONOP**: Count(Non)>0
3. **POS**: Count(Pos)>0 and Count(Neg)=0
4. **NEG**: Count(Neg)>0 and Count(Pos)=0
5. **NEU**: Count(Pos)=0, Count(Neg)=0 and Count(Neu)>0

Preprocessing

Extract single label for each word

1. **NOT**: Count(Not)>0
2. **NONOP**: Count(Non)>0
3. **POS**: Count(Pos)>0 and Count(Neg)=0
4. **NEG**: Count(Neg)>0 and Count(Pos)=0
5. **NEU**: Count(Pos)=0, Count(Neg)=0 and Count(Neu)>0

- Neutral words are dropped since there are only 16 of them
- Words not labeled are also dropped (e.g., Count(Pos)>0 and Count(Neg)>0)

Features

ANTUSD & E-hownet

- CopeOpi score in ANTUSD
- Synonym-Set index (SSI)
  - Concept frame index of a word
  - Each word might belong to many concepts
  - Represented as a binary vector

Features

ANTUSD & E-hownet

- CopeOpi score in ANTUSD
- Synonym-Set index (SSI)
  - Concept frame index of a word
  - Each word might belong to many concepts
  - Represented as a binary vector

Word Embedding

- Corpus: LDC2009T14 (Chinese news)
- Word vectors
- Summation of char vectors

Opinion Extraction

- COP, SSI has lower precision
  - opinion extraction is more semantic-oriented
  - Many words contain single SSI

Feature(s)	Precision	Recall	f-score
COP	0.686	1.000	0.814
SSI	0.693	0.993	0.816
WV	0.784	0.936	0.854
CV	0.765	0.919	0.835
COP+SSI	0.740	0.914	0.818
COP+WV	0.785	0.933	0.853
COP+CV	0.764	0.917	0.833
SSI+WV	0.789	0.937	0.856
SSI+CV	0.772	0.920	0.840
WV+CV	0.808	0.921	0.861

Opinion Extraction

- COP, SSI has lower precision
  - opinion extraction is more semantic-oriented
  - Many words contain single SSI
- Character vectors lead to less precise semantic representation

Feature(s)	Precision	Recall	f-score
COP	0.686	1.000	0.814
SSI	0.693	0.993	0.816
WV	0.784	0.936	0.854
CV	0.765	0.919	0.835
COP+SSI	0.740	0.914	0.818
COP+WV	0.785	0.933	0.853
COP+CV	0.764	0.917	0.833
SSI+WV	0.789	0.937	0.856
SSI+CV	0.772	0.920	0.840
WV+CV	0.808	0.921	0.861

Opinion Extraction

- COP, SSI has lower precision
  - opinion extraction is more semantic-oriented
  - Many words contain single SSI
- Character vectors lead to less precise semantic representation
- Features are complemented; combined features leads to improvement

Feature(s)	Precision	Recall	f-score
COP	0.686	1.000	0.814
SSI	0.693	0.993	0.816
WV	0.784	0.936	0.854
CV	0.765	0.919	0.835
COP+SSI	0.740	0.914	0.818
COP+WV	0.785	0.933	0.853
COP+CV	0.764	0.917	0.833
SSI+WV	0.789	0.937	0.856
SSI+CV	0.772	0.920	0.840
WV+CV	0.808	0.921	0.861

Polarity Classification

- COP leads to a significant better result, reflecting is sentiment-oriented nature

Feature(s)	POS f1	NEG f1	Average f1
COP	0.973	0.976	0.974
SSI	0.792	0.842	0.817
WV	0.870	0.895	0.882
CV	0.829	0.851	0.840
COP+SSI	0.979	0.982	0.980
COP+WV	0.981	0.984	0.982
COP+CV	0.967	0.972	0.970
SSI+WV	0.898	0.915	0.907
SSI+CV	0.868	0.886	0.877
WV+CV	0.899	0.916	0.908

Polarity Classification

- COP leads to a significant better result, reflecting is sentiment-oriented nature
- Combining COP & other features still leads to improvement

Feature(s)	POS f1	NEG f1	Average f1
COP	0.973	0.976	0.974
SSI	0.792	0.842	0.817
WV	0.870	0.895	0.882
CV	0.829	0.851	0.840
COP+SSI	0.979	0.982	0.980
COP+WV	0.981	0.984	0.982
COP+CV	0.967	0.972	0.970
SSI+WV	0.898	0.915	0.907
SSI+CV	0.868	0.886	0.877
WV+CV	0.899	0.916	0.908

Polarity Classification

- COP leads to a significant better result, reflecting is sentiment-oriented nature
- Combining COP & other features still leads to improvement
- Combining word vectors and SSI also leads to improvement

Feature(s)	POS f1	NEG f1	Average f1
COP	0.973	0.976	0.974
SSI	0.792	0.842	0.817
WV	0.870	0.895	0.882
CV	0.829	0.851	0.840
COP+SSI	0.979	0.982	0.980
COP+WV	0.981	0.984	0.982
COP+CV	0.967	0.972	0.970
SSI+WV	0.898	0.915	0.907
SSI+CV	0.868	0.886	0.877
WV+CV	0.899	0.916	0.908

Combined Task

- COP outperforms the others

Feature(s)	Precision	Recall	f-score
COP	0.912	0.927	0.920
SSI	0.706	0.679	0.692
WV	0.737	0.767	0.752
CV	0.689	0.721	0.705
COP+SSI	0.864	0.945	0.903
COP+WV	0.850	0.902	0.875
COP+CV	0.840	0.869	0.854
SSI+WV	0.764	0.796	0.779
SSI+CV	0.732	0.755	0.743
WV+CV	0.764	0.813	0.787

Combined Task

- COP outperforms the others
- Both the numerator of precision and recall are affected by COP's better polarity classification ability
- Only the denominator of precision is affected by COP's worse opinion extraction ability

Precision & Recall

$$P = \frac{correct(opinion) \cap correct(polarity)}{proposed(opinioopinionn)}$$

$$R = \frac{correct(opinion) \cap correct(polarity)}{gold(opinioopinionn)}$$

Feature(s)	Precision	Recall	f-score
COP	0.912	0.927	0.920
SSI	0.706	0.679	0.692
WV	0.737	0.767	0.752
CV	0.689	0.721	0.705
COP+SSI	0.864	0.945	0.903
COP+WV	0.850	0.902	0.875
COP+CV	0.840	0.869	0.854
SSI+WV	0.764	0.796	0.779
SSI+CV	0.732	0.755	0.743
WV+CV	0.764	0.813	0.787

Combined Task

- COP outperforms the others
- Both the numerator of precision and recall are affected by COP's better polarity classification ability
- Only the denominator of precision is affected by COP's worse opinion extraction ability
- WV+CV outperforms WV due to coverage issue

Feature(s)	Precision	Recall	f-score
COP	0.912	0.927	0.920
SSI	0.706	0.679	0.692
WV	0.737	0.767	0.752
CV	0.689	0.721	0.705
COP+SSI	0.864	0.945	0.903
COP+WV	0.850	0.902	0.875
COP+CV	0.840	0.869	0.854
SSI+WV	0.764	0.796	0.779
SSI+CV	0.732	0.755	0.743
WV+CV	0.764	0.813	0.787

Conclusion

- A so far the largest Chinese sentiment dictionary
- Manually sentiment labels & machine estimated sentiment scores
- Three experiments were conducted to demonstrate the usage of ANTUSD

Q & A